Appendix

Appendix A1 Study characteristics: Calderón, Hertz-Lazarowitz, & Slavin, 1998 (quasi-experimental design)

Characteristic	Description
Study citation	Calderón, M., Hertz-Lazarowitz, R., & Slavin, R. (1998). Effects of Bilingual Cooperative Integrated Reading and Composition on students making the transition from Spanish to English reading. *Elementary School Journal, 99,* 153–165.
Participants	The study included 222 Spanish-speaking English language learners in the second (n = 120) and third (n = 102) grades. The students' primary home language was Spanish. A total of 85 third-grade students (52 in the treatment group, 33 in the comparison group) were posttested in English in reading and language.[1] Three intervention and four comparison schools participated in the two-year study.[2]
Setting	English language learners who participated in the study attended seven elementary schools in the El Paso, Texas school district. Overall, 79% of students in the district were Hispanic and 27% had limited English proficiency. The schools selected for inclusion in the study had the highest rates of poverty and the lowest levels of student achievement among the schools in the district with Spanish-dominant English language learners.
Intervention	*BCIRC* students were assigned to cooperative learning teams consisting of four heterogeneously grouped students (that is, groups contained a mix of high, medium, and low achieving students). *BCIRC* attempts to promote student discussion and dialogue during cooperative learning activities designed to help students develop critical thinking and reading comprehension skills as well as the overall ability to use academic English. Activities include partner reading, recognition of key components of a story, vocabulary development, creative writing, and tasks designed to promote reading comprehension. Teachers model reading strategies—such as making and confirming a prediction—before, during, and after reading. Cooperative groups then apply the demonstrated strategy while attempting to comprehend stories selected from their classroom text. Students were taught for two hours each day. One half-hour of the two-hour instruction included English as a Second Language (ESL) instruction for both intervention and comparison groups. Stories for both intervention and comparison classrooms were selected from the Macmillan *Campanitas de Oro* Spanish basal reading series. By the middle of the second grade, students alternated every two weeks between the Spanish basal and the Macmillan *Transitional Reading Program* basal series in English.
Comparison	The comparison group included four schools matched to the three intervention schools on demographic characteristics and academic ranking within the district. Further, individual classes within the intervention schools were matched with classes in the comparison schools on mean pretest scores. Comparison group students used the Macmillan *Campanitas de Oro* Spanish basal reading series and began to alternate between the Spanish basal and the Macmillan *Transitional Reading Program* basal series in English each day. Students received the same amount of instruction (two hours, including one half-hour of ESL instruction) but used the teachers' editions of the McMillan reading series for guidance rather than the *BCIRC* approach. Overall, teachers in the comparison condition were trained in and used round-robin oral reading and workbook practice activities.
Primary outcomes and measurement	The effects of the intervention on English language learner outcomes were assessed using the Norm-Referenced Assessment Program for Texas (NAPT). Although the Texas Assessment of Academic Skills (TAAS) was used in the study to assess reading outcomes, results are not reported here because the measure was administered in Spanish. (See Appendices A2.1 and A2.2 for a more detailed description of the outcome measures.)
Teacher training	Teachers implementing the intervention received extensive staff development, but more specific information about the training was not provided. Teachers in comparison schools received training related to cooperative learning.

[1] There were 102 third-grade students who were pretested (64 in the treatment group, 38 in the comparison group). This sample loss does not exceed WWC limits for study attrition.

[2] Two cohorts of students participated in the study. However, data from only one of the third-grade English language learner cohorts are reported, because students in the other cohort were not assessed in English.

Appendix A2.1 Outcome measures in the reading achievement domain

Outcome measure	Description
Norm-Referenced Assessment Program for Texas (NAPT)—Reading Scale	English language learner outcomes in reading were assessed using the NAPT at the end of the third grade (as cited in Calderón, Hertz-Lazarowitz, & Slavin, 1998). The NAPT yields scores in reading, writing, mathematics, science, and social studies in English.
Percentage of students who met exit criterion from bilingual education	English language learner outcomes in reading were based on the exit criterion from bilingual education. Students who scored above the 40th percentile on the NAPT reading test met the exit criterion for reading achievement.

Appendix A2.2 Outcome measures in the English language development domain

Outcome measure	Description
Norm-Referenced Assessment Program for Texas (NAPT)—Language Scale	English language learner outcomes in English language development were assessed using the NAPT at the end of the third grade (as cited in Calderón, Hertz-Lazarowitz, & Slavin, 1998). The NAPT yields scores in reading, writing, mathematics, science, and social studies in English.
Percentage of students who met exit criterion from bilingual education	English language learner outcomes in English language development were based on the exit criterion from bilingual education. Students who scored above the 40th percentile on the NAPT language test met the exit criterion for English language development.

Appendix A3.1 Summary of study findings included in the rating for the reading achievement domain[1]

			Author's findings from the study					
			Mean outcome[2] (standard deviation[3])		WWC calculations			
Outcome measure	Study sample	Sample size (students)	*BCIRC* group	Comparison group	Mean difference[4] (*BCIRC* – comparison)	Effect size[5]	Statistical significance[6] (at $\alpha = 0.05$)	Improvement index[7]
Calderón, Hertz-Lazarowitz, & Slavin, 1998 (quasi-experimental design)[8]								
Norm-Referenced Assessment Program for Texas—Reading	Grade 3	85	33.16 (15.44)	23.83 (14.98)	9.33	0.61	ns	+23
Domain average[9] for reading achievement						0.61	ns	+23

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement index. Subgroup findings from the same study are not included in these ratings, but are reported in Appendix A4.1.
2. Adjusted means are reported. Kindergarten English and Spanish scores on the Bilingual Syntax Measure served as a pretest covariate.
3. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Calderón, Hertz-Lazarowitz, and Slavin (1998), a correction for clustering was needed, so the significance levels differ from those reported in the original study.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size.

Appendix A3.2 Summary of study findings included in the rating for the English language development domain[1]

			Author's findings from the study					
			Mean outcome[2] (standard deviation[3])		WWC calculations			
Outcome measure	**Study sample**	**Sample size (students)**	***BCIRC* group**	**Comparison group**	**Mean difference[4] (*BCIRC* – comparison)**	**Effect size[5]**	**Statistical significance[6] (at $\alpha = 0.05$)**	**Improvement index[7]**
				Calderón, Hertz-Lazarowitz, & Slavin, 1998 (quasi-experimental design)[8]				
Norm-Referenced Assessment Program for Texas—Language	Grade 3	85	34.90 (15.69)	30.36 (15.91)	4.54	0.29	ns	+11
Domain average[9] for English language development						0.29	ns	+11

ns = not statistically significant

1. This appendix reports findings considered for the effectiveness rating and the average improvement index. Subgroup findings from the same study are not included in these ratings, but are reported in Appendix A4.2.
2. Scores are normal curve equivalents, and adjusted means were provided by the study authors.
3. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes.
4. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
5. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations. Though it is unclear why students in cohort 1 had either one year or two years of *BCIRC* exposure, the issue of two third-grade subsamples does not influence the effect size calculations or ratings presented in the report.
6. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
7. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
8. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools and for multiple comparisons. For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Calderón, Hertz-Lazarowitz, and Slavin (1998), no corrections for clustering or multiple comparisons were needed.
9. This row provides the study average, which in this instance is also the domain average. The WWC-computed domain average effect size is a simple average rounded to two decimal places. The domain improvement index is calculated from the average effect size. Though it is unclear why students in cohort 1 had either one or two years of *BCIRC* exposure, the issue of two third-grade subsamples does not influence the effect size calculations or ratings presented in the report.

Appendix A4.1 Summary of subgroup findings for the reading achievement domain[1]

			Author's findings from the study		WWC calculations			
			Mean outcome[2] (standard deviation[3])					
Outcome measure	Study sample[4]	Sample size (students)	*BCIRC* group	Comparison group	Mean difference[5] (*BCIRC* – comparison)	Effect size[6]	Statistical significance[7] (at $\alpha = 0.05$)	Improvement index[8]
Calderón, Hertz-Lazarowitz, & Slavin, 1998 (quasi-experimental design)[9]								
Norm-Referenced Assessment Program for Texas—Reading	Grade 3—two years	59	36.83 (nr)	23.83 (14.98)	13.00	0.87	Statistically significant	+31
Norm-Referenced Assessment Program for Texas—Reading	Grade 3—one year	59	28.83 (nr)	23.83 (14.98)	5.00	0.33	ns	+13
Percentage of students who met exit criterion from bilingual education—Reading	Grade 3	118	0.32	0.10	0.22	0.87	Statistically significant	+31

ns = not statistically significant
nr = not reported

1. This appendix presents subgroup findings for measures that fall in the reading achievement domain. Findings for the full sample were used for rating purposes and are presented in Appendix A3.1.
2. Scores are normal curve equivalents, and adjusted means are provided.
3. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes. The study authors did not provide standard deviations for subgroups.
4. "One year" represents students who were in the program for one year, and "two years" represents students who were in the program for two years. The study is unclear about why some third-grade students in cohort 1 had one year of *BCIRC* exposure while other third-grade students in the same cohort had two years of exposure.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
6. The appendix table reports the effect sizes, but the WWC could not confirm the effect sizes because the study did not report standard deviations for the subgroups. The effect sizes reported by the study authors were computed as the difference in adjusted scores on the posttest divided by unadjusted control group standard deviations, which differs from the method that the WWC uses to compute effect sizes. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC could not confirm that the effects of the intervention were statistically significant because the study did not include standard deviations for the subgroups.
8. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Calderón, Hertz-Lazarowitz, and Slavin (1998), a correction for clustering was needed, which did not change the statistical significance of the findings reported by the study author for students who had been involved with the intervention for two years. However, the findings for students who had been involved with the intervention for one year became nonsignificant after correcting for clustering.

Appendix A4.2 Summary of subgroup findings for the English language development domain[1]

			Author's findings from the study		WWC calculations			
			Mean outcome[2] (standard deviation[3])					
Outcome measure	Study sample[4]	Sample size (students)	*BCIRC* group	Comparison group	Mean difference[5] (*BCIRC* – comparison)	Effect size[6]	Statistical significance[7] (at $\alpha = 0.05$)	Improvement index[8]
Calderón, Hertz-Lazarowitz, & Slavin, 1998 (quasi-experimental design)[9]								
Norm-Referenced Assessment Program for Texas—Language	Grade 3—two years	59	36.27 (nr)	30.21 (nr)	6.06	0.38	ns	+15
Norm-Referenced Assessment Program for Texas—Language	Grade 3—one year	59	33.73 (nr)	30.21 (nr)	3.52	0.22	ns	+9
Percentage of students who met exit criterion from bilingual education—Language	Grade 3	118	0.39	0.21	0.18	0.53	ns	+20

ns = not statistically significant
nr = not reported

1. This appendix presents subgroup findings for measures that fall in the English language development domain. Findings for the full sample were used for rating purposes and are presented in Appendix A3.2.
2. Scores are normal curve equivalents, and adjusted means are provided.
3. The standard deviation across all students in each group shows how dispersed the participants' outcomes are: a smaller standard deviation on a given measure would indicate that participants had more similar outcomes. The study authors did not provide standard deviations for subgroups.
4. "One year" represents students who were in the program for one year, and "two years" represents students who were in the program for two years.
5. Positive differences and effect sizes favor the intervention group; negative differences and effect sizes favor the comparison group.
6. The effect sizes were provided by the study authors but could not be confirmed by the WWC because the study did not report standard deviations for the subgroups. The effect sizes reported by the study authors were computed as the difference in adjusted scores on the posttest divided by unadjusted control group standard deviations, which differs from the method that the WWC uses to compute effect sizes. Effect sizes for binary measures (for example, the percentage of students who met exit criterion on the NAPT) were calculated using a log odds ratio with Cox adjustment. For an explanation of the effect size calculation, see Technical Details of WWC-Conducted Computations.
7. Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups.
8. The improvement index represents the difference between the percentile rank of the average student in the intervention condition and that of the average student in the comparison condition. The improvement index can take on values between –50 and +50, with positive numbers denoting results favorable to the intervention group.
9. The level of statistical significance was reported by the study authors or, where necessary, calculated by the WWC to correct for clustering within classrooms or schools (corrections for multiple comparisons were not done for findings not included in the overall intervention rating). For an explanation about the clustering correction, see the WWC Tutorial on Mismatch. See Technical Details of WWC-Conducted Computations for the formulas the WWC used to calculate statistical significance. In the case of Calderón, Hertz-Lazarowitz, and Slavin (1998), a correction for clustering was needed, so the significance levels for students who were involved in the intervention for two years differ from those reported by the study authors. The authors did not report statistically significant findings for students who were involved in the intervention for one year.

Appendix A5.1 *Bilingual Cooperative Integrated Reading and Composition* rating for the reading achievement domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of reading achievement, the WWC rated *BCIRC* as having potentially positive effects. It did not meet the criteria for positive effects because it had only one study. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *BCIRC* was assigned the highest applicable rating.

Rating received

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. *BCIRC* met this criterion because it had substantively important positive findings.

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Met. *BCIRC* met this criterion because the one study reviewed did not show a statistically significant or substantively important negative effect or indeterminate effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. *BCIRC* did not meet this criterion because only one study was reviewed.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. *BCIRC* met this criterion because the one study reviewed did not show statistically significant or substantively important negative effects.

[1] For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effects. The WWC also considers the size of the domain level effects for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.

Appendix A5.2 *Bilingual Cooperative Integrated Reading and Composition* rating for the English language development domain

The WWC rates an intervention's effects for a given outcome domain as positive, potentially positive, mixed, no discernible effects, potentially negative, or negative.[1]

For the outcome domain of English language development, the WWC rated *BCIRC* as having potentially positive effects. It did not meet the criteria for positive effects because it had only one study. The remaining ratings (mixed effects, no discernible effects, potentially negative effects, and negative effects) were not considered because *BCIRC* was assigned the highest applicable rating.

Rating received

Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: At least one study showing a statistically significant or substantively important *positive* effect.

  Met. *BCIRC* met this criterion because it had substantively important positive findings.

- Criterion 2: No studies showing a statistically significant or substantively important *negative* effect and fewer or the same number of studies showing *indeterminate* effects than showing statistically significant or substantively important *positive* effects.

  Met. *BCIRC* met this criterion because the one study reviewed did not show a statistically significant or substantively important negative effect or indeterminate effects.

Other ratings considered

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Criterion 1: Two or more studies showing statistically significant *positive* effects, at least one of which met WWC evidence standards for a strong design.

  Not met. *BCIRC* did not meet this criterion because only one study was reviewed.

- Criterion 2: No studies showing statistically significant or substantively important *negative* effects.

  Met. *BCIRC* met this criterion because the one study reviewed did not show statistically significant or substantively important negative effects.

[1] For rating purposes, the WWC considers the statistical significance of individual outcomes and the domain level effects. The WWC also considers the size of the domain level effects for ratings of potentially positive or potentially negative effects. See the WWC Intervention Rating Scheme for a complete description.